ISSN No : 2454-4221(Print)
ISSN No : 2454-423X(Online)

International Journal of Research in Advanced Computer Science Engineering

A Peer Reviewed Open Access International Journal

www.ijracse.com

PREDICTION OF PARKINSON’S DISEASE USING MACHINE LEARNING

Sri. B.Ramesh,D.Avinash, V.UdayaSri, P.Sowmya, Ch.Sai Sandeep.

Department of Computer Science and Engineering, Aditya Institute of Technology and Management, Tekkali.

Abstract

The second most prevalent neurological condition, Parkinson disease (PD), lowers quality of life and causes significant disability. It is incurable.Although the precise cause for Parkinson's disease remains to be known, it is thought to be brought on by a confluence of genetic and environmental variables. The primary pathogenic aspect of Parkinson's disease is the decline of dopamine-producing neurons in the substantia nigra, a particular area of the brain. About 90% of Parkinson's patients who are impacted have speech issues. The average age of onset is around 70 years, and as people get older, the incidence increases considerably. In this study, we investigated the usage of Decision Tree, Naive Bayes, Extra Tree Classifier, XGBoost, and KNN algorithms using a dataset of speech features to predict Parkinson's disease. The dataset had 756 rows and 755 columns, and we used recursive feature elimination to choose the top 100 features (RFE). Accuracy, precision, recall, and F1-score metrics were used to assess the models' performance.Metrics like accuracy, precision, recall, and F1-score were used to assess the models' performance. Our study demonstrates the potential of ML techniques in developing a reliable and accurate predictive model for Parkinson's disease diagnosis, which could lead to earlier intervention and better disease management. Our findings indicate that, when compared to the other models, the XGBoost model had the highest accuracy, at 91.4%.

Keywords: **Decision Tree, Naive Bayes, Extra Tree Classifier, XGBoost, KNN, Machine Learning, Recursive Feature Elimination.**

1.Introduction

Millions of people worldwide are affected by Parkinson's disease, a chronic and progressive neurological ailment. For effective therapy and illness management, Parkinson's disease must be identified early and accurately predicted. Based on a variety of clinical, genetic, and imaging data, machine learning has emerged as a potent technique for forecasting the start and course of Parkinson's disease. In this research, we seek to create a model using machine learning that, given a collection of input features, can precisely predict a person's risk of having Parkinson's disease. We will investigate different machine learning techniques, including decision trees, random forests, support vector machines, and neural networks. We will assess each algorithm's performance using measures like accuracy, precision, recall, and F1-score. To find the key Parkinson's disease predictors, we will also use feature extraction and dimensionality reduction. Our ultimate objective is to create a predictive model using machine learning that is trustworthy and accurate and that can help doctors diagnose and treat Parkinson's disease early on, enhancing patient outcomes as well as life quality.

Cite this article as: Sri. B.Ramesh,D.Avinash, V.UdayaSri, P.Sowmya, & Ch.Sai Sandeep, "Prediction Of Parkinson’s Disease Using Machine Learning", International Journal of Research in Advanced Computer Science Engineering, (IJRACSE), Volume 8 Issue 11, April 2023, Page 30-38.

1.1 Causes of parkinsons disease

Parkinson's disease is a complex and multi factorial disorder that results from the loss of dopamine-producing neurons in the brain. The exact causes of Parkinson's disease are not fully understood, but there are several factors that are believed to contribute to the development of the disease.

- Genetic Factors: Genetic mutations have been identified as a significant risk factor for Parkinson's disease. Several genes have been associated with an increased risk of Parkinson's disease, including SNCA, LRRK2, and PARK2. However, it's important to note that most cases of Parkinson's disease are not directly caused by genetic mutations and have a complex interplay between genetic and environmental factors.
- Environmental Factors: Exposure to environmental toxins has been linked to an increased risk of Parkinson's disease. Exposure to pesticides, herbicides, and other chemicals have been associated with an increased risk of Parkinson's disease. In addition, exposure to metals such as lead and manganese have also been linked to Parkinson's disease.
- Age: Parkinson's disease is more common in older adults, and age is considered to be a significant risk factor for the disease.
- Gender: Men are more likely to develop Parkinson's disease than women, although the reasons for this are not fully understood.
- Head Trauma: Repeated head injuries and trauma have been associated with an increased risk of Parkinson's disease.
- Other Medical Conditions: Several medical conditions have been linked to an increased risk of Parkinson's disease, including diabetes, depression, and REM sleep behavior disorder.

1.2 Signs and symptoms of parkinsons disease

Parkinson's disease is a degenerative neurological disorder that primarily affects movement, causing symptoms such as tremors, stiffness, and difficulty with balance and coordination. The signs and symptoms of Parkinson's disease can vary from person to person, but some common ones include:

- Tremors or shaking of the hands, arms, legs, jaw, or face, especially at rest.
- Stiffness or rigidity of the muscles, making it difficult to move or perform daily activities.
- Bradykinesia, or slowness of movement, which can result in difficulty initiating or completing movements, as well as difficulty with fine motor tasks.
- Postural instability or difficulty maintaining balance, which can lead to falls.
- Changes in gait or walking pattern, such as shuffling or a stooped posture.
- Reduced facial expression or "masking," which can make it difficult to convey emotions.
- Changes in speech, such as softness, slurring, or hesitation.
- Loss of smell, constipation, or other non-motor symptoms.

Literature survey

Several research works have been done on Parkison's disease prediction. Different techniques are been implemented from machine learning and deep learning for Parkison's disease prediction & achieved different results for different methods.

Timothy J. Wroge [1] Because of its underlying cognitive and neuromuscular function, biomarkers extracted from human voice can provide insight into neurological illnesses like Parkinson's disease (PD). A million Americans are thought to have PD, a progressive neurodegenerative condition, and each year, roughly 60,000 new clinical cases are identified. In his paper, he investigates how well supervised classification techniques, like deep neural networks, may be used to precisely classify patients with the illness. The machine learning models' peak accuracy of 85% is higher than the average clinical diagnosis accuracy of non-experts (73.8%) and the average accuracy of movement disorder are (79.6% without follow-up, 83.9% after follow-up).

Ramadugu Akhil [2] Parkinson's disease is a neurological ailment brought on by damage to our body's nerve cells. Parkinson's disease is caused by a lack of dopamine production, or, to put it another way, an 80% reduction in the number of cells that make dopamine. He applied machine learning to determine whether a patient had Parkinson's disease or not. They made use of the vocal characteristics of numerous patients, both those with and without Parkinson's disease diagnoses. They employed many machine learning techniques, including KNN, Random Forest, and Logistic Regression. In order to improve accuracy while employing machine learning models, PCA is used to simplify the data and make it easier to read with the least amount of data loss. They obtained a greater SVM accuracy of 91.8%.

Harshvardhan Tiwari [3]Parkinson's disease (PD) is a progressive neurological disorder that kills the substantia nigra cells that produce dopamine (di-ortho-phenyl-alanine). No reliable test can consistently distinguish PD from other illnesses with comparable clinical manifestations. The history and physical examination serve as the foundation for the clinical diagnosis. In order to determine whether a person has Parkinson's disease or not based on the voice input parameters, six classification algorithms were used: Logistic Regression, Support vector machine (SVM), Decision tree, K-Nearest Neighbour (KNN), and XGBOOST (Extreme gradient boosting). Extreme gradient boosting, or XGBOOST, has the highest accuracy rating of 90.8%, making it superior to all other methods.

Mousumy Kundu[4] Neurodegeneration may be delayed and disability may be avoided with an early and accurate diagnosis of PD and access to effective treatment. The current method of diagnosis, such as a brain scan, is an expensive procedure. In some research, voice recorded data and machine learning algorithms were used to distinguish PD patients from healthy people. The dataset is based on speech recording data for PD status prediction from the machine learning repository at the University of California, Irvine (UCI ML). They have suggested an improved method of data pre-processing that increases the precision of PD diagnosis predictions. AdaBoost, MARS, and SVM are employed as algorithms.

Aditi Govindu[5] A neurological condition called Parkinson's disease (PD) affects 60% of persons over 50. Parkinson's disease (PWP) patients struggle with speech impairment and movement issues, which makes it difficult for them to travel for appointments for treatment and monitoring. Early discovery of PD enables treatment, allowing patients to live normal lives. During the training of four ML models, research was done using the MDVP audio data of 30 PWP and healthy individuals. When Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression model classification results are compared, Random Forest classifier is found to be the best Machine Learning (ML) technique for PD identification.

Senturk[6]When dopamine-producing brain cells are disrupted, Parkinson's disease develops. Dopamine is a chemical that enables brain cells to connect with one another. Parkinson's is a neurodegenerative brain disease that progresses gradually. Neurodegenerative conditions result in the death of brain cells. Dopamine enables humans to move with grace and harmony. Parkinson's disease (PD) motor symptoms emerge when 60–80% of dopamine-producing cells are destroyed because there is insufficient dopamine available for the body to use. Three algorithms—CART, SVM, and ANN—were used in this study, and the accuracy values were 90.76%, 93.84%, and 91.54%, respectively. For the early identification of Parkinson's disease, an FS-based decision support system was created employing the voice signal characteristics of both PD patients and healthy individuals.The findings show that combining FS approaches with classification methods is extremely beneficial, especially when working with speech data, which might yield hundreds of phonetic variables. With the use of the proposed early diagnosis technique, PD can be accurately identified in its early stages and the worsening of the disease's symptoms can be halted.

Celik [7]Parkinson's disease is a particular illness brought on by the death of dopamine-producing brain cells. In order to predict Parkinson's disease, various classification techniques, including Logistic Regression, Support Vector Machine, Extra Trees, Gradient Boosting, and Random Forest, are compared in this study. The classification step used a total of 1208 speech data sets made up of 26 features collected from Parkinson's patients and non-patients. The use of correlation maps broadens the dataset's feature space. These correlation maps are created utilising features that were acquired through the use of Principal Component Analysis (PCA), Information Gain (IG), and all features, respectively. The best result came from the Linear Kernel Support Vector Machine (75.49%).

Almeida[8] They looked at vocal signal processing for Parkinson's disease detection. The method assesses the usage of four machine learning methods and eighteen feature extraction approaches to categorise data generated from sustained phonation and speech activities. Speech is related to the pronunciation of a brief sentence in Lithuanian, and phonation is related to the vowel /a/ voicing challenge. Two microphone channels from an acoustic cardioid microphone (AC) and a smartphone (SP) were used to capture the audio tasks, allowing performance from various microphone types to be compared. Equal Error Rate (EER) and Area Under Curve (AUC) metrics from Detection Error Tradeoff (DET) and Receiver Operating Characteristic curves, as well as Accuracy, Specifiability, and Sensitivity, were used to examine the classifaction performance.We contrast this strategy with other strategies that make use of the same collection of data. They demonstrate that phonation activities were more effective than speech tasks at detecting illness. The accuracy, AUC, and EER values for the AC channel's best performance were 94.55%, 0.87, and 19.01%, respectively. They have obtained accuracy of 92.94%, AUC of 0.92, and EER of 14.15 % using the SP channel.

Moshkova[9]For the purpose of identifying Parkinson's disease, they used the kinematic parameters of hand movements (PD). Leap Motion sensors were used to record the hand motions of 16 PD patients (N16) and 16 members of the control group (N16). Based on MDS-UPDRS part 3, the following three motor tasks were selected: finger tapping (FT), pronation-supination of the hand (PS), and opening-closing hand movements (OC). 25

kinematic characteristics for the signal from the sensor were derived using key points. Using a specifically created user application, the key point determination was done utilising the maximums and minimums finding algorithm as well as manual marking. Several feature extraction methods were employed for the binary classification (PD or non-PD), for each motor task separately and for the three combined.The following four classifiers were trained: kNN, SVM, Decision Tree (DT), and Random Forest (RF). In the 8-fold cross-validation mode, testing was done. The combination of the most important traits of both hands produced the best results. The outcomes were as follows for each task: FT 95.3%, OC 90.6%, and PS 93.8%.

Balaji[10] Neurologists frequently use a number of clinical symptoms to diagnose Parkinson's disease (PD) and assign a severity level using the Unified Parkinson Disease Rating Scale (UPDRS). A gait classification method based on machine learning that can help the clinician identify the stages of Parkinson's disease is presented. Gait pattern, which is important for evaluating human mobility, is an important biomarker to identify if an individual has Parkinson's disease (PD) or is in good condition.Thus, we use the vertical ground reaction force (VGRF) gait dataset and use statistical analysis to extract the minimal feature vector. For statistical and kinematic analyses that predict the severity of PD, four supervised machine learning algorithms decision tree (DT), support vector machine (SVM), ensemble classifier (EC), and Bayes classifier (BC)are used.

3. Dataset

The dataset used for this study was obtained from the Kaggle Repository, which is accessible to the general public. This is a thorough summary of the dataset:

Sl. No.	Dataset Name	Sources	Attribute Type	Number of Attributes	Number of Instances
1	PD speech features dataset	Kaggle Repository	Continuous and Binary	755	756

Table 1: Details of Database

4. Proposed Methodologies

Figure 1 shows the steps in the proposed workflow which involves the pre-processing of data, training, and testing with specified models, evaluation of results and prediction of Parkinson’s disease using machine learning. This work is implemented in Python 3.

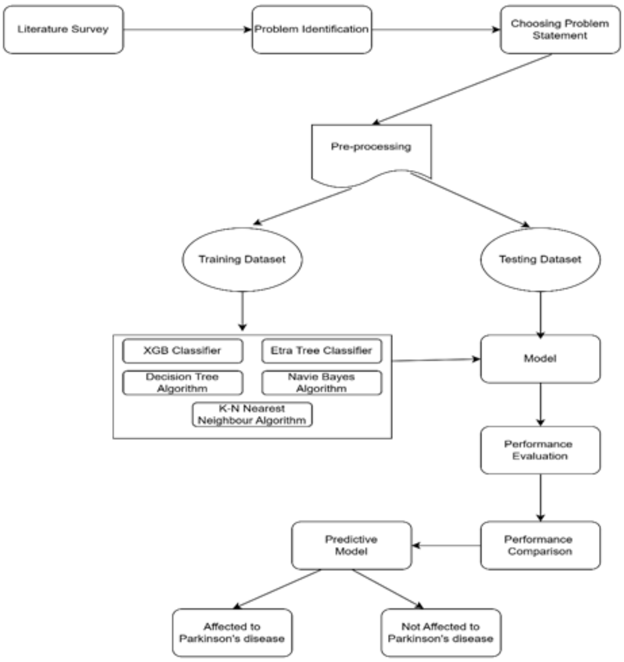


Figure 1.Steps in the proposed parkinsons detection solution

4.1. Data Pre-Processing

Pre-processing involves converting raw data into a format that is meaningful and comprehensible. Real-world data frequently has errors and null values, which causes it to be inconsistent and incomplete. A good outcome is always produced by well-preprocessed data. To deal with incomplete and inconsistent data, a variety of pre-processing techniques are utilised, including handling missing values, outlier detection, data discretization, data reduction (dimension and numerosity reduction), etc. Imputation has been used to solve the dataset's concerns with missing values.

4.2. Training and Testing

Model The whole dataset has been split into two parts i.e. one part is training dataset and the other one is testing dataset with a ratio of 80:20 respectively. Figure 2 shows the final training, testing and validation sets on which classification has been performed.

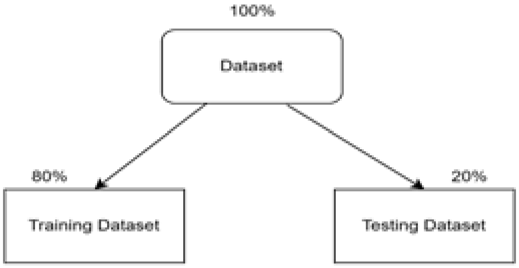


Figure 2.Training and Testing dataset

4.2.1 XGBoost Classifier

Extreme Gradient Boosting, or XGBoost, is a gradient boosting method built for speed and efficiency.Iteratively adding weak learners to a model, each learner attempting to fix the mistakes of the preceding learner, is how the method operates. To avoid overfitting, XGBoost employs a regularised objective function and supports both categorical and continuous data. To enhance efficiency and speed up calculation,the technique also allows parallel processing, tree pruning, and early termination. For classification, regression, and ranking problems, XGBoost is extensively utilised in a variety of industries, including finance, healthcare, and computer vision.

4.2.2 K-Nearest Neighbour

The supervised machine learning method KNN (K-Nearest Neighbors) is utilised for both classification and regression applications. The algorithm predicts the output based on the majority or average of the k nearest neighbour outputs after locating the k closest instances to a new instance in the training data. KNN is a straightforward, non-parametric model that works with both continuous and categorical data. The algorithm's performance can be impacted by the selection of distance measure and the number of neighbours (k). In many different industries, including bioinformatics, image recognition, and recommendation systems, KNN is frequently employed.

4.2.3 Decision Tree Algorithm

Decision Trees are a supervised learning method that may be applied to classification and regression issues, however it is typically chosen for dealing with classification issues. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. The Decision Node and Leaf Node are the two nodes of a decision tree. Whereas Leaf nodes are the results of those decisions and do not have any further branches, Decision nodes are utilized for making any decision and have many branches.

4.2.4 NAIVE BAYES(NB)Algorithm

The Nave Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set. The Naive Bayes Classifier is among the most straightforward and efficient classification algorithms available today. It aids in the development of quick machine learning models capable of making accurate predictions. Being a probabilistic classifier, it makes predictions based on the likelihood of an object.

4.2.5 Extra Tree Classifier

Very Randomized Trees Classifier, also known as Extra Trees Classifier, is a form of aggregation

learning technique that synthesises the findings of various de-correlated decision trees gathered in a "forest" to provide its classification outcome. The only way it differs conceptually from a Random Forest Classifier is in how the decision trees in the forest are built. The initial training sample is used to build each decision tree in the Extra Trees Forest. Finally, each tree is given a random sample of k features from the feature-set at each test node, from which it must choose the best feature to divide the data according to certain mathematical criterion. There are numerous de-correlated decision trees produced as a result of this random sampling of features.

5. Results and Discussion

Using the classification report and confusion matrix, the outcome is evaluated in terms of Accuracy, F1-Score, Precision Score, and Recall Score. How accurately the model is trained will determine the outcome.

To determine how effectively a categorization model is achieving a goal, performance measurement is essential. A classification model's efficiency and performance on the test dataset are assessed using performance assessment metrics. The right measures, such as the confusion matrix, accuracy, specificity, and sensitivity, should be used to assess the performance of the model. The performance measures are determined using the formulas below:

$$\text{Precision Score} = \frac{TP}{(TP+FP)}$$

$$\text{Recall Score} = \frac{TP}{(TP+FN)}$$

$$\text{F1- Score} = \frac{TP}{TP+1/2(FP+FN)}$$

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)}$$

5.1 Performance Evaluation metrics

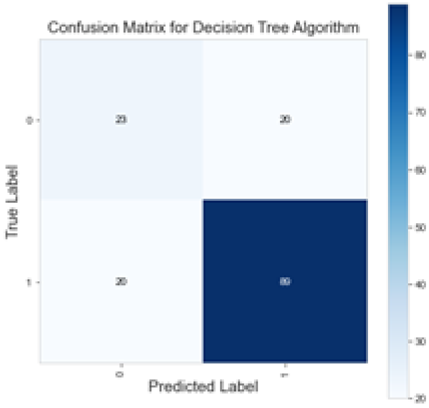
It's important to measure performance to see how well a classification model accomplishes a goal. On the test dataset, the classification model's performance is assessed using performance evaluation measures. It is crucial to select the appropriate metrics to assess the success of the model, such as the confusion matrix.accuracy, sensitivity, and so on. The performance measurements are calculated using the following formulas:

With the Parkinson's data, experimental results of several machine learning approaches with all features selected have been shown. We first applied all 5 models on 755 features and determined the models' precision, F1-Score, recall, and accuracy. Some of the 755 features were redundant, so we used recursive feature elimination (RFE) and ranking algorithms on all of them before choosing 100 features to build the best models, which we then measured for precision, F1-Score, recall, and accuracy. Below is a detailed breakdown of the overall performance metrics for each machine learning classifier using the as fore mentioned dataset.

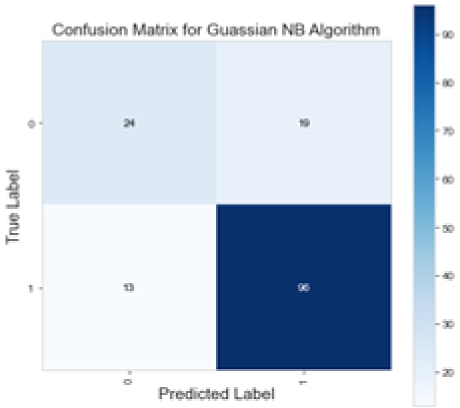
Classifier	Precision Score	Recall Score	F1-Score	Training Accuracy	Testing Accuracy
Decision Tree	81.81	82.56	82.19	100	74.34
Guassian Naive Bayes	83.47	88.07	85.71	81.4	78.94
Extra Tree	88.98	96.33	92.51	100	88.15
XGBoost	92.8	95.4	94.11	100	91.44
KNN	81.96	91.74	86.58	86.92	79.60

Table 2: Overall Results for Parkinsons detection using Machine Learning Algorithms

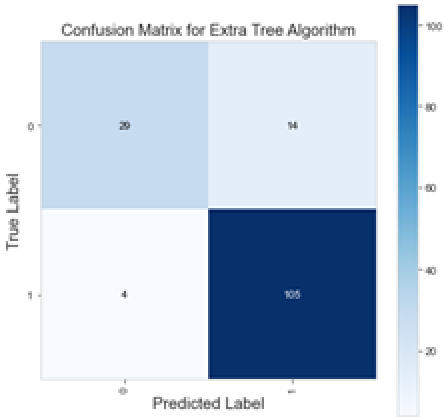
On the Parkinson's disease dataset, evaluation of various machine learning models revealed accuracy ranging from (78% to 91%) on the original dataset. For the original dataset, decision tree classifier provided the lowest accuracy (78%), guassian nb produced the next-highest accuracy (79%), K-neighbour produced the next-highest accuracy (80%), extra tree produced the greatest accuracy (87%), and xgboost produced the highest accuracy (91%). The outcomes of the prediction model are also described by the confusion matrices of all Machine Learning algorithms



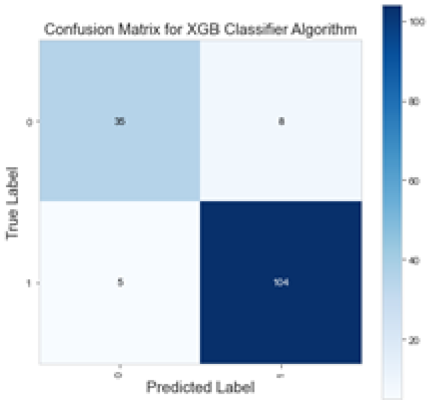
a)



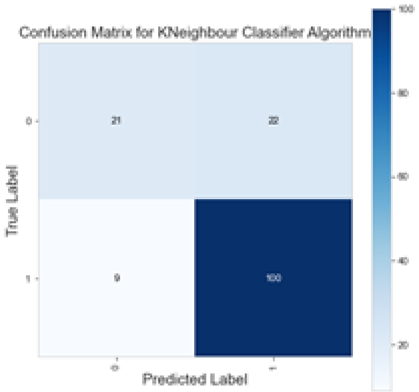
b)



c)



d)



e)

Figure 3: Confusion Matrices for Different Machine Learning Algorithms

6. Conclusion

In this study, a variety of machine learning approaches were used to try and detect Parkinson's disease. To evaluate the performance of the models used to detect Parkinson's disease on the Parkinson's Dataset, a variety of performance evaluation indicators were employed. When XGBoost's classifier was put up against all other machine learning models, it exhibited high performance with 91% accuracy

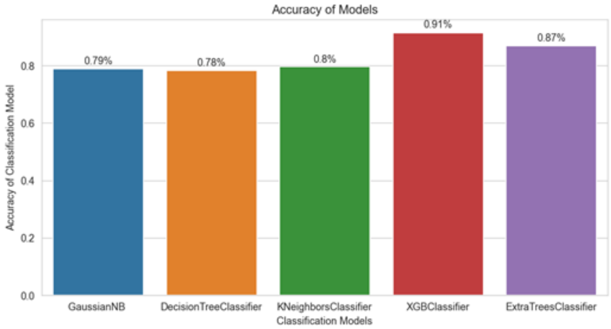


Figure 4: Testing accuracy comparison of Various Machine Learning Algorithms

7. References

[1] Wroge, Timothy J., et al. "Parkinson’s disease diagnosis using machine learning and voice." 2018 IEEE signal processing in medicine and biology symposium (SPMB). IEEE, 2018.

[2] Akhil, Ramadugu, Mohammed Rayyan Irbaz, and M. Aruna. "Prediction of Parkinson’s Disease Using Machine Learning." Annals of the Romanian Society for Cell Biology (2021): 5360-5367.

[3] Tiwari, Harshvardhan, et al. "Early prediction of parkinson disease using machine learning and deep learning approaches." EasyChair Preprint 4889 (2021): 1-14.

[4] Kundu, Mousumy, et al. "An optimized machine learning approach for predicting Parkinson’s disease." Int. J. Mod. Educ. Comput. Sci.(IJMECS) 13.4 (2021): 68-74.

[5]Aditi Govindu, Sushila Palwe,Early detection of Parkinson's disease using machine learning,Procedia Computer Science,Volume 218,2023.

[6] Senturk, Zehra Karapinar. "Early diagnosis of Parkinson’s disease using machine learning algorithms." Medical hypotheses 138 (2020): 109603.

[7] Celik, Enes, and Sevinc Ilhan Omurca. "Improving Parkinson's disease diagnosis with machine learning methods." 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). Ieee, 2019.

[8] Almeida, Jefferson S., et al. "Detecting Parkinson’s disease with sustained phonation and speech signals using machine learning techniques." Pattern Recognition Letters 125 (2019): 55-62.

[9] Moshkova, Anastasia, et al. "Parkinson’s disease detection by using machine learning algorithms and hand movement signal from LeapMotion sensor." 2020 26th Conference of Open Innovations Association (FRUCT). IEEE, 2020.

[10] Balaji, E., D. Brindha, and R. Balakrishnan. "Supervised machine learning based gait classification system for early detection and stage classification of Parkinson’s disease." Applied Soft Computing 94 (2020): 106494.